KU LEUVEN

FACULTY OF ECONOMICS
AND BUSINESS

Understanding recommendation quality using embeddings

Reusens M, Haegemans T, Lemahieu W, Snoeck M, Baesens B, Sels L.

KBI_1712

Understanding recommendation quality using embeddings

Reusens, Michael Haegemans, Tom Lemahieu, Wilfried
Snoeck, Monique Baesens, Bart
Sels, Luc
firstname.lastname@kuleuven.be

August 1, 2017

Abstract

Current methods of evaluating the quality of recommender systems are based on averages of metrics such as the average normalized discounted cumulative gain, average diversity and average reciprocity. Averages of metrics give a good sense of the overall quality of the recommendations, but not of how their quality is distributed with respect to the recommendation system's users or items. This paper presents a visual method, based on embedding a high dimensional content feature-space into a 2D image, that is capable of providing insights in which users are receiving high quality recommendations and how biased recommendation quality is with respect to different types of users. Through a proof of concept in the domain of job recommendation we show that our method allows business people to come to relevant answers to the question "For which of my users does my recommender system work well/poorly?".

1 Introduction

Organisations often offer recommendation services to reduce information asymmetries between themselves and their customers. Such a service can be implemented by a recommender system that suggests items matching the individual preferences of its users. When a recommendation is tailored to the preferences of a user, it will help him/her to make a more informed and more swift decision about a certain choice s/he faces (e.g. which movie to watch, which item to buy next, . . .) (Dahlman, 1979). In many cases, when the users can decide purchasing a good or service faster and more informed, the organisation is likely to gain an increase in revenue and customer satisfaction (Schafer, Konstan, & Riedl, 1999).

Unfortunately, sometimes, recommender systems provide recommendations that are not tailored to the preferences of their users, i.e. some users might receive low quality recommendations. When the users receive too many low quality recommendations, they might lose trust in the overall system, or the recommendation provider as a whole (O'Donovan & Smyth, 2005). Consequently, the organisation that provides the recommendation service might injure a decrease in revenue and customer satisfaction.

To ensure that the users of recommendation systems receive high quality recommendations, organisations aim to optimise the performance of these systems. Such an optimisation is carried out by the developers of the system in close cooperation with several business stakeholders such as decision makers and marketeers. Typically, first, the developers and business stakeholders interpret information about the system's performance, such as metrics and indicators, and next, based on this information, they decide whether and how the recommender system should be improved. As such, it is important that the information about the recommender system's performance provides insights about when low quality recommendations occur and that this information can be interpreted by both developers and business stakeholders.

In the literature and in current practice, information about the performance of a recommender system is usually presented by metrics that average the performance of the recommendations over the set of all users. Some commonly reported average performance metrics are, for example, the average- root mean squared error (RMSE), normalized discounted cumulative gain (NDGC) and diversity (Herlocker, Konstan, Terveen, & Riedl, 2004). These average metrics and indicators provide a good feel for the quality of the recommender system's performance and provide an easy way of comparing multiple recommender systems against each other

However, average performance metrics and indicators are not very useful when developers or business users want to gain deeper understanding about low quality recommendations. On the one hand, when the performance of individual recommendations is averaged, important information about these individual recommendations is lost. For example, based on the average performance, it is impossible to know whether there were many minor irrelevant recommendations or only a few severely irrelevant recommendations. Second, average performance indicators do not relate the performance of the system to the user the recommendation was presented to. For example, developers and business users might ask themselves whether the low quality recommendations occur randomly or systematically? And, if these low quality recommendations occur systematically, they typically want to understand for which users the low quality recommendations occur.

Therefore, in this work we present a methodology that allows the inspection of the distribution of low quality recommendations with respect to a set of features that are of interest by developers and business stakeholders. The proposed approach allows inspection of the errors in an intuitive way by both parties, which will lead to a better understanding of low quality recommendations.

2 Related Work

In recent years a strand of research has emerged that aims to understand and model the situations in which recommender systems perform poorly. The first of such models comes from (Bellogín, 2011). They acknowledge that performance prediction is already an established research topic in information retrieval (IR). The authors show that by translating the concept of query clarity from IR to the recommender systems domain, a recommender system's performance averaged over all users can be predicted. (Kille, 2012) presents the concept of recommendation difficulty on a user level. (Matuszyk & Spiliopoulou, 2014) quantifies the difficulty of a recommendation dataset. They also present a method of visually assessing the eligibility of collaborative filtering by looking at a heat map showing the number of co-rated items between each user-user pair. The authors of (Gras, Brun, & Boyer, 2015) aim to identify users that have an atypical rating pattern. They show that users with atypical rating patterns will more likely receive poor recommendations from collaborative filtering recommender systems. In (Ekstrand & Riedl, 2012), the authors show that different recommender algorithms make different errors, meaning that in a lot of cases when a specific algorithm fails there is usually another algorithm that succeeds. They also present a model based on user- and item meta features (no content features) that predicts if a given recommender system will make a good or bad interest prediction for a given user-item pair.

Interpreting complex data by visual inspection is something that has proven its value in marketing research. (Seret, Verbraken, Versailles, & Baesens, 2012; Azcarraga, Hsieh, Pan, & Setiono, 2005) present a methodology based on self-organising maps (SOM) that allows marketeers to get a 2D visual representation of their whole user base.

Our research contributes to the studies presented above by combining the goal of understanding when recommender systems fail with visualisation techniques that allow business users to gain insights in the user population. Furthermore, we are interested in looking at users from a business perspective (described by content features) and not from an algorithm perspective (described by rating counts, rating variations, number of near neighbours, . . .).

3 Embedding High Dimensional Data to 2D Images

Embedding is the translation of data from an origin-dimension D_x to a target-dimension D_y, with x and y the sizes of respective dimensions. When the target-dimension is smaller than the origin-dimension, this is also called dimensionality reduction (Roweis & Saul, 2000). The use of embeddings in this study is to translate high-dimensional content data, describing users of a recommender system, to a 2-dimensional coördinate system used to create an image that reveals groups and structure within the users of said recommender system. Translating a set of high dimensional observations $N_{orig} \in D_{orig}$ to a set of points on an image $N_2 \in D_2$ requires an embedding $f(N_{orig} \rightarrow N_2)$. In order for the resulting image to be interpretable, pairwise distances in N_{orig} should be preserved as well as possible in N_2.

Because of the decrease in dimensions, information will almost always be lost, and a perfect preservation of pairwise distances will not be possible. Differences between different embedding techniques mainly originate from choices of what pieces of information to retain. For example, the t-sne embedding (Maaten & Hinton, 2008) used in the experiments presented in Section 5 focuses on maintaining small pairwise distances, but fails to maintain larger pairwise distances.

Embeddings have been used to visualise and interpret various high dimensional datasets, such as for visualizing groups in handwritten digit images (Hinton & Roweis, 2002), grouping images of similar objects, and grouping similar customers in e-commerce settings (Seret et al., 2012). In the latter example it was shown that by embedding high-dimensional customer data to a 2D representation, non-technical business users were enabled to reason about the different types of users present in their database.

In this work, we build further upon the idea of using embeddings to reason about a business' user base. In the next section, we propose to use embeddings to get an easier overview of the recommendation quality distribution over all users.

4 Proposed Approach

This section presents our approach, which aims to answer the following questions that often arise when developing recommender systems in a business:

1. Is the recommendation error distributed randomly, or higher (lower) in certain groups of users?

2. How can we describe the groups of users that receive better (worse) recommendations?

4.1 Step 1: Feature Selection

A key aspect of answering these questions is being able to express types of users: e.g.: young people, men, highly educated, ... Some organisations might even already have such understanding of users in place (e.g. from an earlier clustering/marketing exercise). In this case they can select the same variables that constitute that understanding to generate the embedding. Using features familiar within the business will likely make it easier to communicate when a recommender system works well/poorly.

4.2 Step 2: Generate Embedding

Starting from the features selected in the previous step, one should generate a 2D embedding as discussed in Section 3. Depending on the embedding technique chosen, this step includes the tuning of parameters until all the users are layed out on the image so that you can explain different sections of the embedding using colourings. An example of such colourings can be seen in the three left images of Figure 1. Be aware that a good embedding is one from which groups of users can be visually separated. Finding the best (or simply a good) embedding for a specific dataset can be a process of trial and error (van der Maaten, 2017). Some embedding strategies will even lead to different results for the same parametrisation, making it even harder to provide strict guidelines that will lead to an insightful embedding.

4.3 Step 3: Define Recommender Quality Metric(s)

Our proposed technique works independently from the system designer's choice of recommender algorithm and evaluation metrics. Once a suitable algorithm and evaluation metric is selected, the recommender system should be evaluated for a random set of users. This results in a dataset with for each user the quality measured by one or more evaluation metrics.

4.4 Step 4: Compare Recommender Error Colouring with Origin Space Feature Colouring

The embedding can now be coloured with each recommendation quality metric selected in Step 3. An example of such a colouring can be found in the rightmost image of Figure 1. Now visual inspection can be performed on this image in combination with earlier colourings of the embedding (such as the three leftmost images in Figure 1).

To get an idea of the answer to the first question, merely looking at the quality-colouring will give an indication whether recommendation quality is higher/lower in specific areas of the embedding (meaning there is probably a bias in quality towards a specific user group) or appears to be randomly distributed over the image (meaning there is likely no bias in quality towards a specific user group).

To answer the second question, one should compare the quality-colouring with the feature-colourings. If feature-coloured zones can be found that overlap with concentrations of high/low recommendation quality in the quality-colouring, this is an indication that the recommender algorithm is positively/negatively biased towards this type of user.

5 Application: Inspecting Job Recommendation Quality at the Flemish Public Employment Services

We applied our methodology to job recommendation using data provided by the Flemish public employment services (PES). This use case is extremely relevant for our methodology since the question "for which of our job seekers/vacancies does the recommender system work well?" is important for public organisations that aim to provide equal services to everyone.

For a random sample of 10,000 job seekers with at least 1 vacancy click top-10 job recommendations were generated using user-user collaborative filtering with Jaccard distance metric. A rating of 1 is given to job seeker-vacancy pairs when the job seeker clicked on the vacancy, and NA if he/she did not. Recommendation quality was measured using recall@10 on a leave-10-out test set. For each of the 10,000 job seekers the last 10 ratings were left out. The choice of recommender system and evaluation metric is kept basic for demonstration purposes, but could be substituted with any other algorithm and metric if looking for the best-in-class job recommender system.

A 2D embedding of the 10,000 sample job seekers was created. The origin dimension consisted of 14 variables commonly used by the PES to distinguish between job seekers in their daily business activities: age, sex, location, education level, native language, being a new user (registered with the PES for less than 10 days) and others. As embedding algorithm we used the Barnes-Hut implementation of t-sne (with perplexity = 250 and maximum number of iterations = 2000), implemented in the *Rtsne* package for R (Maaten & Hinton, 2008; van der Maaten, 2014; Krijthe, 2015). Parameter selection was performed manually, using visual inspection of the embedding quality. t-sne was chosen over alternatives, such as principal component analysis (PCA) or stochastic neighbourhood embedding (sne) (Hinton & Roweis, 2002), because it has a smaller tendency to group all observations in the centre of the image, resulting in a better visual separation of groups that may exist in the data (Maaten & Hinton, 2008).

Figure 1 shows four colourings of the the resulting embedding. The three left colourings are example overlays of origin dimension features (the other origin feature colourings have been left out of this paper) and the right figure shows the recall@10 of the recommender system overlaid on the embedding.

By visual inspection, we can come to interesting insights about the job seeker data, and recommender quality distribution.

Not all origin features are separated equally well. For example, the job seekers with low education level are separated almost perfectly from those with a high education level. New users, while still grouped together, are more scattered throughout the embedding. This is a common observations in many embeddings, since some variables will be more discriminative than others. The quality of an embedding is highly dependent on distance metric and parametrisation. Even very similar parametrisation can lead to totally different embedding results. Coming up with a good embedding can be a process of trial and error.

The recall@10 colouring definitely does not look random. There are clear zones in the embedding where the recall equals 0, and zones where the recall is strictly larger than 0. We can already suspect a bias in recall w.r.t. the origin dimensions as the quality does not look randomly distributed.

Next, the zones of higher/lower recall are compared with patterns in the origin feature colourings. We do not notice any obvious similarity with the colouring based on education level alone, but similarities in colouring can be

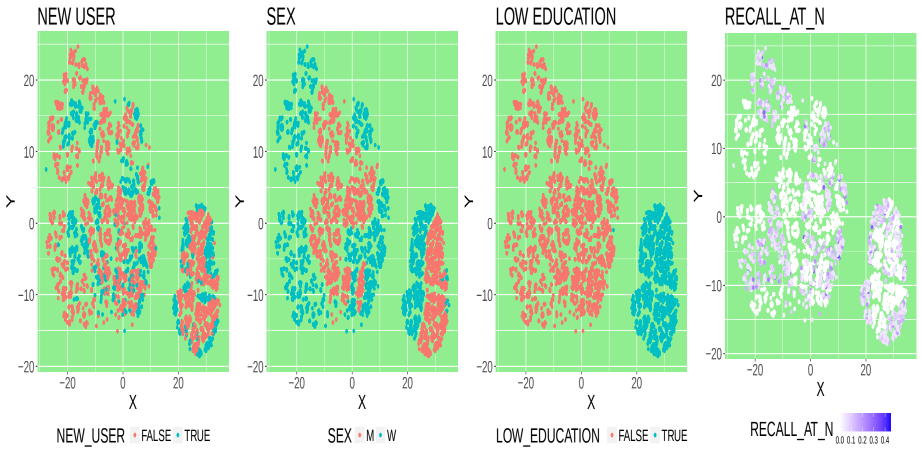

Figure 1: t-sne embedding of 10,000 random job seekers coloured with the following features (from left to right): New User (T/F), Sex (M/W), Low education level (T/F), recall@10 of User-User collaborative filtering ([0,1])

seen with the colouring for new users with low education (on the right side of the embedding), and for the colouring based on sex with high education (on the left side of the embedding). For the feature colourings not included in this paper, no strong colouring similarities could be detected. This leads us to believe that the recommender system employed in our experiments is negatively biased in terms of recall for older users with a low education and men whith a not-low education.

In order to verify our insights drawn from the embedding, a decision tree was trained on all origin features to predict if the recall@10 would be 0 (TRUE), or not (FALSE). The resulting tree can be seen in Figure 2. The decision tree leads to similar, but not the same, conclusions. It confirms our visual insight that there is a bias related to sex and being a new user or not, but not the link we visually observed with education.

6 Discussion

Section 5 showed that the proposed methodology can lead to interesting, but not perfectly accurate, insights in the distribution of recommendation quality over its users. We see this methodology as just one piece in the larger unsolved puzzle of successfully evaluating and understanding recommender systems. We identify the following valuable research directions that both address limitations in our methodology and build further upon it.

The first problem we will address in future work is the lack of certainty in conclusions drawn by visual inspection: how can we make sure that our eyes do not deceive us into seeing patterns in randomness, or vice versa, fail to see

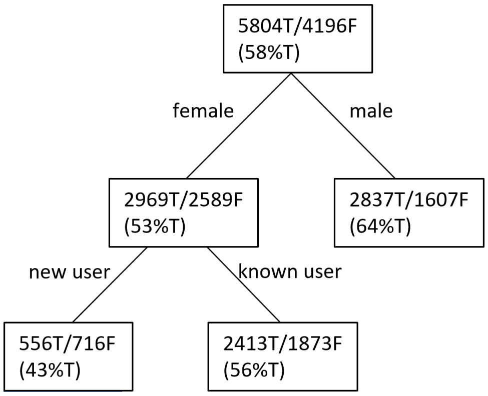

Figure 2: Decision tree trained on origin features with target label T if recall@10 was 0 and F if it was larger than 0.

existing patterns because of poor embedding quality. In the use case we approached this by building a predictive model on the same data and compared conclusions that come out of it with conclusions drawn from visual inspection of the embedding. We also see potential in looking at statistical tests on metrics such as information gain and correlation between the origin dimensions and recommendation quality. Visual inspection and statistical approaches seem complementary techniques, as they each have their benefits. Visualisations are more easily interpreted by non-technical users, and once an embedding is created, it can be used to understand various other aspects about a user base. On the other hand, statistical metrics are less likely to lead to imperfect conclusions, when interpreted correctly. Further research on the interoperability between both methods is in order.

User testing on interpretability of visualisations is another key next step. We only informally checked with employees of the Flemish PES whether they could come to correct conclusions about recommendation quality on their own. A more elaborate user study on which visualisation results in the best user conclusions is key in further developing this research track.

Our methodology can be analogously applied to inspecting recommendation quality for items. This could help businesses answer questions such as: Which items are (not) being successfully recommended? A more complicated extension to our methodology could look into generating embeddings that allow for insights into combinations of users and items: Which user-item combinations our recommender system is predicting to be relevant (do not) work?

We also see application domains for this visualisation methodology beyond recommender systems. Our method also seems appropriate to gain insights in the quality of predictive modelling and root causes of data quality.

7 Conclusion

In this work we presented a methodology that allows developers and business stakeholders to visually inspect the distribution recommendation quality with respect to a set of features that are of interest. The core idea of our approach is to first create a two dimensional embedding of the multidimensional data using embedding techniques such as t-sne, and next to project the recommendation quality on this embedding. As such, it becomes easy for recommender system's developers or business stakeholders to visually inspect the quality of the recommended users (or items).

Acknowledgements

This research was funded by the career management analytics research chair, sponsored by the Flemish PES (VDAB) and by the data quality chair, sponsored by KBC Group NV.

References

Azcarraga, A. P., Hsieh, M., Pan, S. L., & Setiono, R. (2005). Extracting salient dimensions for automatic SOM labeling. *IEEE Trans. Systems, Man, and Cybernetics, Part C*, *35*(4), 595–600. Retrieved from http://dx.doi.org/10.1109/TSMCC.2004.843177 doi: 10.1109/TSMCC.2004.843177

Bellogín, A. (2011). Predicting performance in recommender systems. In *Proceedings of the 2011 ACM conference on recommender systems, recsys 2011, chicago, il, usa, october 23-27, 2011* (pp. 371–374). Retrieved from http://doi.acm.org/10.1145/2043932.2044009 doi: 10.1145/2043932.2044009

Dahlman, C. J. (1979). The Problem of Externality. *Journal of Law and Economics*, *22*(1), 141 – 162.

Ekstrand, M. D., & Riedl, J. (2012). When recommenders fail: predicting recommender failure for algorithm selection and combination. In *Sixth ACM conference on recommender systems, recsys '12, dublin, ireland, september 9-13, 2012* (pp. 233–236). Retrieved from http://doi.acm.org/10.1145/2365952.2366002 doi: 10.1145/2365952.2366002

Gras, B., Brun, A., & Boyer, A. (2015). Identifying users with atypical preferences to anticipate inaccurate recommendations. In *WEBIST 2015 - proceedings of the 11th international conference on web information systems and technologies, lisbon, portugal, 20-22 may, 2015* (pp. 381–389). Retrieved from http://dx.doi.org/10.5220/0005412703810389 doi: 10.5220/0005412703810389

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004, jan). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. Retrieved from http://dl.acm.org/citation.cfm?id=963770.963772 doi: 10.1145/963770.963772

Hinton, G. E., & Roweis, S. T. (2002). Stochastic neighbor embedding. In *Advances in neural information processing systems 15 [neural information processing systems, NIPS 2002, december 9-14, 2002, vancouver, british columbia, canada]* (pp. 833–840). Retrieved from http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding

Kille, B. (2012). Modeling difficulty in recommender systems. In *Proceedings of the workshop on recommendation utility evaluation: Beyond rmse, RUE 2012, dublin, ireland, september 9, 2012* (pp. 30–32). Retrieved from http://ceur-ws.org/Vol-910/paper7.pdf

Krijthe, J. H. (2015). Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation [Computer software manual]. Retrieved from https://github.com/jkrijthe/Rtsne (R package version 0.11)

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Matuszyk, P., & Spiliopoulou, M. (2014). Predicting the performance of collaborative filtering algorithms. In *4th international conference on web intelligence, mining and semantics (WIMS 14), WIMS '14, thessaloniki, greece, june 2-4, 2014* (pp. 38:1–38:6). Retrieved from http://doi.acm.org/10.1145/2611040.2611054 doi: 10.1145/2611040.2611054

O'Donovan, J., & Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference on intelligent user interfaces, IUI 2005, san diego, california, usa, january 10-13, 2005* (pp. 167–174). Retrieved from http://doi.acm.org/10.1145/1040830.1040870 doi: 10.1145/1040830.1040870

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323–2326.

Schafer, J. B., Konstan, J. A., & Riedl, J. (1999). Recommender systems in e-commerce. In *EC* (pp. 158–166). Retrieved from http://doi.acm.org/10.1145/336992.337035 doi: 10.1145/336992.337035

Seret, A., Verbraken, T., Versailles, S., & Baesens, B. (2012). A new som-based method for profile generation: Theory and an application in direct marketing. *European Journal of Operational Research*, *220*(1), 199–209. Retrieved from http://dx.doi.org/10.1016/j.ejor.2012.01.044 doi: 10.1016/j.ejor.2012.01.044

van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, *15*(1), 3221–3245. Retrieved from http://dl.acm.org/citation.cfm?id=2697068

van der Maaten, L. (2017). *t-SNE.* https://lvdmaaten.github.io/tsne/. ([Online; accessed 27-March-2017])

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be